Theorem 1 outlines the smoothing properties of stochastic gradient descent (SGD) by positing that the gradient's total variance remains constant across iterations. This variance relates to the gradient size and the Lipschitz constant's limited rate. It provides a quantitative method to evaluate SGD stability, highlighting the importance of gradient magnitude and Lipschitz continuity for the algorithm's smooth performance. Additionally, review the momentum update rule as shown in (20) and (21).

$$v_{t+1} = \beta v_t + \alpha \nabla f_s\left(\theta_t\right) \tag{1}$$

$$\theta_{t+1} = \theta_t - v_{t+1} \tag{2}$$

For the expectation of the gradient, it follows that $E\left[\nabla f_s(\theta_t)\right] = E[g_t]$. Hence, the derivation of the momentum's expectation can be expressed as (22).

$$E\left[v_{t+1}\right] = \alpha \sum_{i=0}^{t} \beta^i E\left[g_{t+1-i}\right] \tag{3}$$

The Lipschitz condition for the gradient:

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L\|\theta_1 - \theta_2\| \tag{4}$$

where L is the Lipschitz constant. This equation gives an upper bound on how much the gradient can change based on the distance between parameters θ_1 and θ_2.

The expected gradient difference is bounded by:

$$\mathbb{E}[\|\nabla f(\theta_{t+1}) - \nabla f(\theta_t)\|] \leq L\mathbb{E}[\|\theta_{t+1} - \theta_t\|] \tag{5}$$

which shows that the Lipschitz constant defines the upper bound for the expected gradient difference, but does not guarantee the actual difference unless other conditions are met.

The variance of the gradient can further refine this bound:

$$\text{Var}(\nabla f(\theta_t)) \leq \frac{L}{n} \sum_{i=1}^{n} \|\theta_{t+1}^i - \theta_t^i\| \tag{6}$$

where n is the number of parameters, and this formula expresses how the gradient variance is influenced by both the Lipschitz constant and the parameter distance.

Theorem 2: a lower limit on the expected gradient of the accumulation time T for a given momentum approach is established.

$$\begin{gathered}\sum_{t=1}^{T} \log\left(E\left[L_t\right]\right) \leq \\ \sum_{t=1}^{T} \log\left(\left[E\left\|\triangledown f\left(\theta_{T+1}\right) - \triangledown f\left(\theta_1\right)\right\|\right]\right) \\ \wedge - \log(\alpha \sum_{i=0}^{t} \beta^i E\left[\|g_{t+1-i}\|\right]\end{gathered} \tag{7}$$

Momentum imposes a logarithmic limit on loss expectation L_k, combining current and past gradients with weights. This integration of historical gradient data enhances momentum's ability to avoid local minima and saddle points. Historical gradients influence not only parameter updates but also the direction of these updates. This smoothing effect helps overcome challenges from noisy or unstable learning conditions. Rapid convergence to optimal solutions ensures accurate sentiment analysis, which is crucial for reliable health-related Q&A systems, thereby fostering patient trust and satisfaction.

Thus, momentum usually leads to lower, more favorable limits on L_k, speeding up convergence and yielding better outcomes in complex contexts. In the Adam method, parameter adjustments rely on both the first moment (mean) and the second moment (unbiased variance approximation) of the gradient. An estimate of the stochastic gradient is computed using the mean $\widehat{m}_t = E\left[g_t\right]$ and the unbiased variance $\widehat{v}_t = E^2\left[g_t\right]$, as demostrated in (27).

$$\theta_{t+1} = \theta_t - \alpha \frac{E\left[g_t\right]}{\sqrt{E^2\left[g_t\right] + \mathit{Var}\left(g_t\right)}} \tag{8}$$

Theorem 3: The lower bound of $E\left[L_t\right]$ is obtained according to Adam's method.

$$\begin{gathered}\sum_{t=1}^{T} \log\left(E\left[L_t\right]\right) \leq \\ \sum_{t=1}^{T} (\log\left(E\left[\|\nabla f\left(\theta_{T+1}\right) - \triangledown f\left(\theta_1\right)\|\right]\right) \\ - \log\left(\alpha \frac{E\left[\|g_t\|\right]}{\sqrt{E^2\left[\|g_t\|\right] + \mathit{Var}\left(\|g_t\|\right)}}\right))\end{gathered} \tag{9}$$

The Adam method's dynamic traits may yield clearer minima than other algorithms, risking solution robustness. By modifying gradient data, the model enhances its ability to gauge nuanced shifts in patient sentiments during training, leading to improved understanding and prediction of concerns. Additionally, the gradient's variance significantly affects the size and direction of Adam's updates, with excessive variance risking instability.

Theorem 4: introduces g_t as a stochastic gradient defined by $g_t = \nabla f_s(\theta_t)$. When the variance $var(g_t)$ of the stochastic variable g_t converges towards the Cramer-Rao lower limit C.

$$\lim_{t \to \infty} var\left(g_t\right) = C \tag{10}$$

Equation (29) implies that as t progresses, g_t will approach its unbiased expected value $E\left[g_t\right]$, as denoted in (30).

$$\lim_{t \to \infty} P\left(g_t - E\left[g_t\right] = 0\right) = 1 \tag{11}$$